DOCUMENT RESUME

ED 458 812 FL 026 968

AUTHOR Nakamura, Yuji
TITLE Rasch Measurement and Item Banking: Theory and Practice.
PUB DATE 2001-10-00
NOTE 15p.; Supported in part by a research grant from Tokoya Keizai University.
CONTRACT CPU04-00
PUB TYPE Reports - Research (143)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Item Response Theory; *Second Language Instruction; Second Language Learning; Student Evaluation; *Test Items; Test Validity; Testing
IDENTIFIERS Rasch Model

ABSTRACT

The Rasch Model is an item response theory, one parameter model developed that states that the probability of a correct response on a test is a function of the difficulty of the item and the ability of the candidate. Item banking is useful for language testing. The Rasch Model provides estimates of item difficulties that are meaningful, irrespective of the ability level tested. This paper focuses mainly on how the model can contribute to the feasibility of item banking in terms of language testing. The present research deals mainly with the following basic aspects of item banking: calibration of items for storage; the measurement of student abilities; and the advantages and limitations of item banking. Item characteristics can be determined either by traditional item statistics (Classical Test Theory) or a newer method of estimating item statistics called Item Response Theory. In this paper, Item Response Theory is used for dealing with item characteristics. The sample data for the research in this paper were taken from a multiple choice test that consisted of 10 items and was given to 105 students. It is concluded that Item Response Theory facilitates item banking by allowing all of the items to be calibrated and positioned on the same latent continuum by means of a common metric, and allows additional items to be added subsequently without the need to locate and retest the original sample of examinees. Furthermore, an item bank permits the construction of tests of known reliability and validity based on appropriate selection of item subsets from the bank without further need for trial in the field. (Contains 12 references.) (KFT)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 458 812

Rasch Measurement and Item Banking: Theory and Practice

Yuji Nakamura

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Yuki Nakamura

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

コミュニケーション科学　第15号　所載抜刷
(2001年10月)

BEST COPY AVAILABLE

Rasch Measurement and Item Banking : Theory and Practice

Yuji Nakamura

1. Introduction

Henning (1989) says that latent trait measurement or item response theory refers primarily, but not entirely, to three families of analytical procedures. These three are identified as the one-parameter (or Rasch Model), the two-parameter, and the three-parameter logistic models. The first parameter is a scale of person ability and item difficulty ; the second parameter is a continuous estimate of discriminability ; the third parameter is an index of guessing (Henning, 1989).

The Rasch model is an item response theory (IRT), one-parameter modél developed by George Rasch, which states that the probability of a correct response is a function of the difficulty of the item and the ability of the candidate. The term oneparameter refers to the item difficulty parameter (Davies et al, 2000). The model makes it possible to predict the likelihood of a correct answer to a given test item on the basis of the knowledge of two variables : item difficulty and person ability.

Oscarson (1999) claims that application of the Rasch model provides the researcher with information on how to organize the test items in terms of level of difficulty, spread of item difficulty, test length, etc. in order to obtain optimal precision of measurement. This can be viewed as the general and primary function of this model.

Among the applications of the Rasch model, item banking is useful for language testing. Item banking is the process of creating a pool of items with known and invariant measurement characteristics. The Rasch model provides estimates of item difficulties that are meaningful, irrespective of ability level tested. This paper focuses mainly on how the model can contribute to the feasibility of item banking in terms of language testing.

2. Purpose of the research and research design

The present research deals mainly with the following basic aspects of item banking : 1) the calibration of items for storage, 2) the measurement of students' abilities, and 3) the advantages and limitations of item banking.

Item characteristics can be determined either by traditional item statistics (called Classical Test Theory) or a newer method of estimating item statistics called Item Response Theory. In this paper Item Response Theory is used for dealing with item characteristics. The sample data for the present research was taken from a multiple-choice test that consisted of 10 items and was given to 105 students (cf. Appendix 1).

3. Theoretical background and rationale

An item bank, according to Beeston (2000), is a large collection of test items that have been classified and stored in a database so that they can be retrieved at a later time and chosen for new tests. The items are all classified according to certain characteristics such as the topic of a text, the testing point of an item or statistical information about item difficulty. It is important for the difficulty level of each item to be determined on a common scale of difficulty so that any combination of items can be put into a new test and the item difficulties added together to give a precise measure of the difficulty of that test.

Gronlund (1998) also says that item banks are files of various suitable test items and, further, that they are coded by subject area, instructional level, instructional objective, and various pertinent item characteristics (e.g., item difficulty and discriminating power). Item banks are commonly used 1) for the construction of equivalent or alternate forms of standardized tests (different combinations of homogeneous items are drawn from the bank), and 2) as the basis for computer adaptive tests (items at a suitable level of difficulty for individual candidates are retrieved from the computer bank as required).

Choppin (1979) describes an item bank as a large collection of test questions organized and catalogued like the books in a library. The idea is that the test user can select test items as required to make up a particular test. Since one would think in terms of item banks with several thousand items, the number of possible tests which could be composed from such a bank is huge. Choppin claims that the great advantage of this system is its flexibility. Tests can be long or short, hard or difficult, as the teacher desires.

According to Davies et al (1999), the requirements for an item bank are 1) an adequate pool of test items, 2) an inventory of the abilities and content that each item purports to measure, 3) statistical data indicating the characteristics of each item as evidenced in test trailing (e.g., item difficulty and item discrimination indices), and 3) a theory or construct of ability that enables the meaning of scores on any test that may be constructed from the banked items to be interpreted. Davies et al further suggest that latent trait models are particularly useful in item banking because they have the advantage of allowing item scores to be translated into estimates of ability on a common scale. Thus, all tests deriving from a logit scale item bank are automatically equated since a person's score on any combination of test items can be converted into an ability estimate on the common bank scale. This means that any group of people can be given a test made up of items particularly suitable for them, yet all the results can be compared to one another.

Hozayin (2000) proposes three important characteristics of item banks : 1) storage, 2) coding and 3) item characteristics (difficulty and discrimination). Firstly, in the phase of storage, item banks are stored in a computer file, in files especially designed for this purpose. Secondly, in the phase of coding, items are coded according to their content : by subject area, by instructional level and by instructional objective. Lastly, the phase of item characteristics can be seen in terms of item difficulty and item discrimination. Item difficulty is a familiar concept in educational testing. It shows how many test takers got an item correct. Item discrimination refers to the ability of the test item to distinguish between those students who have learned the material and those who have not (Hozayin, 2000).

Wright and Bell (1985) claim that the definition of an item bank goes beyond storage and coding. An item bank is not just a collection of items but a bank of test items that are carefully calibrated. To calibrate items means to standardize them and make them more precise. In the process of increasing precision, we need to investigate item characteristics (item difficulty and item discrimination) mentioned above.

When items are calibrated and joined in a common bank, any cluster of them can be used to measure ability on the same scale as ability measured by any other cluster of these items. This is called test-free person measurement. In other words, because items have been calibrated for difficulty it is possible to select items to match the known ability range of the examinees.

4. Practical procedures of item calibration and person measurement through Rasch calibration for item banking

Rasch calibration applies a probabilistic model to data in order to construct linear measures. In addition to being linear, these measures are accompanied by relevant estimates of their statistical validity and precision. This greatly enhances our information concerning the measure of the test-takers and the calibration of the items.

It is impossible to estimate a finite ability for persons who correctly answer all or none of a set of items. In such cases all we know is that these persons are more (or less) able than this test can measure. Thus, the first step in calibration involves setting aside persons with extreme scores (cf. Bode and Wright 1999).

Let us take a look at our sample data in Table 1 below. In this table 15 students out of 105 should be set aside because of extremely good scores (in this case they all got 10 items correct). (See Appendix 2 for some terms in the table).

Table 1 STUDENT STATISTICS : MEASURE ORDER

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	REALSE	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	SCORE CORR.	Students
25	10	10	87.8	18.7	MAXIMUM		ESTIMATED		MEASURE	S 25
26	10	10	87.8	18.7	MAXIMUM		ESTIMATED		MEASURE	S 26
27	10	10	87.8	18.7	MAXIMUM		ESTIMATED		MEASURE	S 27
28	10	10	87.8	18.7	MAXIMUM		ESTIMATED		MEASURE	S 28
37	10	10	87.8	18.7	MAXIMUM		ESTIMATED		MEASURE	S 37
44	10	10	87.8	18.7	MAXIMUM		ESTIMATED		MEASURE	S 44
45	10	10	87.8	18.7	MAXIMUM		ESTIMATED		MEASURE	S 45
46	10	10	87.8	18.7	MAXIMUM		ESTIMATED		MEASURE	S 46
73	10	10	87.8	18.7	MAXIMUM		ESTIMATED		MEASURE	S 73
74	10	10	87.8	18.7	MAXIMUM		ESTIMATED		MEASURE	S 74
79	10	10	87.8	18.7	MAXIMUM		ESTIMATED		MEASURE	S 79
86	10	10	87.8	18.7	MAXIMUM		ESTIMATED		MEASURE	S 86
89	10	10	87.8	18.7	MAXIMUM		ESTIMATED		MEASURE	S 89
94	10	10	87.8	18.7	MAXIMUM		ESTIMATED		MEASURE	S 94
95	10	10	87.8	18.7	MAXIMUM		ESTIMATED		MEASURE	S 95
11	9	10	74.5	10.9	.71	.4	.34	-.8	.61	S 11
22	9	10	74.5	10.9	.71	.4	.34	-.8	.61	S 22
30	9	10	74.5	11.1	1.05	.1	.75	-.2	.24	S 30
31	9	10	74.5	12.3	1.28	.3	2.70	1.0	.33	S 31
34	9	10	74.5	11.1	1.05	.1	.75	-.2	.24	S 34
36	9	10	74.5	11.1	1.05	.1	.75	-.2	.24	S 36
40	9	10	74.5	11.1	1.05	.1	.75	-.2	.24	S 40

43	9	10	74.5	10.9	.71	-.4	.34	-.8	.61	S 43
47	9	10	74.5	11.1	1.05	.1	.75	.2	.24	S 47
48	9	10	74.5	10.9	.71	-.4	.34	-.8	.61	S 48
53	9	10	74.5	12.0	1.22	.3	1.62	.4	.11	S 53
66	9	10	74.5	10.9	.71	-.4	.34	-.8	.61	S 66
70	9	10	74.5	10.9	.71	-.4	.34	-.8	.61	S 70
72	9	10	74.5	10.9	.71	-.4	.34	-.8	.61	S 72
76	9	10	74.5	10.9	.71	-.4	.34	-.8	.61	S 76
80	9	10	74.5	11.1	1.05	.1	.75	-.2	.24	S 80
85	9	10	74.5	10.9	.95	-.1	.56	-.4	.38	S 85
88	9	10	74.5	10.9	.71	-.4	.34	-.8	.61	S 88
92	9	10	74.5	10.9	.95	-.1	.56	-.4	.38	S 92
93	9	10	74.5	10.9	.95	-.1	.56	-.4	.38	S 93
97	9	10	74.5	10.9	.95	-.1	.56	-.4	.38	S 97
99	9	10	74.5	10.9	.71	-.4	.34	-.8	.61	S 99
102	9	10	74.5	10.9	.71	-.4	.34	-.8	.61	Sa 2
104	9	10	74.5	12.1	1.23	.3	1.75	.5	-.14	Sa 4
105	9	10	74.5	12.0	1.22	.3	1.62	.4	-.11	Sa 5
3	8	10	65.7	8.8	1.11	.2	1.04	.1	.20	S 03
4	8	10	65.7	8.3	.73	-.6	.53	-.8	.64	S 04
6	8	10	65.7	8.3	.83	-.4	.69	-.5	.51	S 06
8	8	10	65.7	9.6	1.32	.6	1.34	.4	-.05	S 08
13	8	10	65.7	8.3	.83	-.4	.69	-.5	.51	S 13
20	8	10	65.7	8.3	.95	-.1	1.09	.1	.30	S 20
21	8	10	65.7	8.4	1.03	.1	.81	-.3	.33	S 21
29	8	10	65.7	9.2	1.22	.4	1.18	.2	.08	S 29
32	8	10	65.7	8.3	.95	-.1	1.09	.1	.30	S 32
33	8	10	65.7	8.8	1.12	.3	.90	-.1	.23	S 33
35	8	10	65.7	8.3	.95	-.1	1.09	.1	.30	S 35
38	8	10	65.7	8.8	1.12	.3	.90	-.1	.23	S 38
39	8	10	65.7	8.8	1.12	.3	.90	-.1	.23	S 39
41	8	10	65.7	8.3	.73	-.6	.53	-.8	.64	S 41
49	8	10	65.7	10.5	1.59	1.1	3.39	2.2	-.69	S 49
50	8	10	65.7	9.2	1.23	.5	1.06	.1	.10	S 50
54	8	10	65.7	8.3	.63	-.9	.44	-1.0	.74	S 54
62	8	10	65.7	8.3	.73	-.6	.53	-.8	.64	S 62
63	8	10	65.7	10.0	1.44	.8	1.74	.9	-.26	S 63
67	8	10	65.7	9.6	1.33	.6	2.56	1.6	-.26	S 67
68	8	10	65.7	8.3	.63	-.9	.44	-1.0	.74	S 68
69	8	10	65.7	8.3	.63	-.9	.44	-1.0	.74	S 69
77	8	10	65.7	8.3	.73	-.6	.53	-.8	.64	S 77
87	8	10	65.7	8.3	.91	-.2	.91	-.1	.38	S 87
2	7	10	59.7	8.1	1.23	.6	1.28	.6	.08	S 02
7	7	10	59.7	7.4	.81	-.6	.77	-.5	.56	S 07
10	7	10	59.7	8.2	1.24	.7	1.13	.3	.11	S 10

24	7	10	59.7	9.5	1.67	1.7	2.38	2.2	-.55	S 24
42	7	10	59.7	9.4	1.62	1.6	2.23	2.0	-.48	S 42
57	7	10	59.7	8.0	1.18	.5	1.33	.6	.10	S 57
64	7	10	59.7	7.4	.88	-.4	.81	-.4	.49	S 64
65	7	10	59.7	8.3	1.27	.8	1.27	.5	.04	S 65
71	7	10	59.7	7.8	1.12	.4	1.18	.4	.19	S 71
78	7	10	59.7	7.4	.81	-.6	.77	-.5	.56	S 78
82	7	10	59.7	9.2	1.55	1.4	1.691	.2	-.30	S 82
83	7	10	59.7	8.2	1.23	.7	1.10	.2	.13	S 83
84	7	10	59.7	7.4	.81	-.6	.77	-.5	.56	S 84
91	7	10	59.7	7.4	.60	-1.4	.50	-1.3	.81	S 91
96	7	10	59.7	7.4	.70	-1.0	.60	-1.0	.69	S 96
101	7	10	59.7	7.4	.70	-1.0	.60	-1.0	.69	Sa 1
103	7	10	59.7	7.4	.79	-.7	.67	-.8	.60	Sa 3
1	6	10	54.6	7.9	1.31	1.1	1.32	.9	.01	S 01
5	6	10	54.6	7.4	1.13	.5	1.08	.3	.23	S 05
9	6	10	54.6	7.1	1.04	.2	.98	-.1	.33	S 09
12	6	10	54.6	7.3	1.11	.4	1.03	.1	.27	S 12
15	6	10	54.6	6.9	.72	-1.2	.66	-1.2	.69	S 15
19	6	10	54.6	7.7	1.24	.9	1.20	.6	.10	S 19
23	6	10	54.6	6.9	.62	-1.7	.57	-1.6	.80	S 23
51	6	10	54.6	6.9	.72	-1.2	.66	-1.2	.69	S 51
56	6	10	54.6	7.6	1.21	.8	1.24	.7	.11	S 56
60	6	10	54.6	7.9	1.31	1.1	1.32	.9	.01	S 60
81	6	10	54.6	6.9	.83	-.7	.77	-.8	.56	S 81
98	6	10	54.6	8.4	1.48	1.6	1.73	1.9	-.24	S 98
100	6	10	54.6	7.7	1.23	.8	1.26	.7	.09	Sa 0
14	5	10	49.9	7.0	1.07	.3	1.02	.1	.30	S 14
16	5	10	49.9	6.8	.71	-1.4	.67	-1.4	.70	S 16
17	5	10	49.9	6.8	.81	-.9	.76	-1.0	.59	S 17
18	5	10	49.9	6.8	.94	-.3	.88	-.4	.45	S 18
52	5	10	49.9	6.8	.94	-.3	.88	-.4	.45	S 52
55	5	10	49.9	6.8	.75	-1.2	.70	-1.2	.65	S 55
58	5	10	49.9	7.5	1.22	.9	1.31	1.0	.09	S 58
59	5	10	49.9	6.8	.60	-2.1	.56	-1.9	.82	S 59
75	5	10	49.9	6.9	1.02	.1	1.08	.3	.32	S 75
90	5	10	49.9	7.8	1.31	1.3	1.38	1.2	.00	S 90
61	3	10	40.3	7.9	1.17	.5	1.14	.3	.15	S 61
MEAN	7.	10.	63.4	8.9	.99	.0	.96	-.1		
S. D.	1.	0.	8.7	1.6	.26	.8	.56	.8		

Similarly, it is impossible to estimate a finite difficulty for items that are answered correctly by all (or none) of the persons taking them. Then all we know is that these items are too easy or too difficult for this sample of persons. Data editing also sets aside

items with extreme scores (cf. Bode and Wright 1999).

Again let us look at our sample data in Table 2 below. In this table there are no items that should be set aside because of extreme scores.

Table 2 ITEMS STATISTICS: MEASURE ORDER

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	REALSE	INFIT		OUTFIT		SCORE CORR.	Items
					MNSQ	ZSTD	MNSQ	ZSTD		
2	42	90	65.0	2.5	1.21	2.4	1.22	2.1	.16	-I 0002
9	53	90	59.2	2.3	.87	-1.5	.87	-1.3	.52	I 0009
3	59	90	55.9	2.7	1.23	2.2	1.31	2.1	.12	I 0003
1	66	90	51.7	2.6	.81	-1.7	.69	-1.9	.57	I 0001
5	66	90	51.7	2.6	.92	-.7	.78	-1.3	.47	I 0005
8	72	90	47.4	2.8	.94	-.4	.97	-.1	.38	I 0008
4	73	90	46.6	2.9	1.04	.2	.97	-.1	.29	I 0004
6	75	90	44.9	3.0	.94	-.3	.82	-.7	.38	I 0006
10	78	90	42.0	3.4	1.09	.4	1.06	.2	.20	I 0010
7	83	90	35.6	4.1	.98	-.1	.92	-.2	.23	I 0007
MEAN	67.	90.	50.0	2.9	1.00	.0	.96	-.1		
S. D.	12.	0.	8.2	.5	.13	1.3	.18	1.3		

Since cases with extreme scores have been removed (15 in persons, and none in items in the sample data), the data for the remaining persons and items are used in the following analysis.

In order to free these persons and item scores from sample size and test length, they are transformed into proportions of their maximum possible values. To linearize these proportions, they are converted to log odds, or logits (usually from -3 to 3), by taking the natural log of the proportion incorrect for items or failures for persons. This transforms the proportions to a linear scale (Bode and Wright 1999).

Logit scores (person ability and item difficulty) are further transformed, in the present research, into measure scores on a 0-100 scale, which should be more familiar to readers and which should make the test data easier to understand. Also we can avoid negative scores of low achievers and easy items.

Accordingly, Table 1 above shows item-free person ability measures, while Table 2 above shows person-free item difficulty measures. For example, in Table 1, Student 11 has a Rasch ability measure of 74.5 (although there are other students with the same measure in these data), which is an estimate of this person's ability, regardless of which items he responded to. Another example is Student 61, who has a Rasch ability measure

of 40.3, and this is the lowest ability among the 95 measured students.

As for person-free item measures, in Table 2, Item 2 has an item calibration or difficulty measure of 65.0, which is an estimate of this item's difficulty, regardless of the ability level of the persons who responded to it. Another example is Item 7, which has an item calibration or difficulty of 35.6, and this is the easiest item among these 10 items.

In addition, Rasch analysis provides two estimates of misfit : infit and outfit. Infit is sensitive to irregular patterns of responses for items close to a person's ability level. Outfit is sensitive to unexpected responses to items far from the person's ability level. Both are useful indicators of potential problems. Large outfit indicates the presence in the data of unexpected off-target responses. Large infit, in contrast, indicates a central pattern of response incoherence. Although overfit or small misfit values provide insight into how an item set might be shortened by deleting redundant items, they are generally not a concern (Bode and Wright 1999). Therefore, we can be entirely flexible about misfit.

Let us examine Table 3 for the present sample test data. This table shows the 10 calibrated items in the misfit order.

Table 3 ITEMS STATISTICS : MISFIT ORDER

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	REALSE	INFIT		OUTFIT		SCORE	Items
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	
3	59	90	55.9	2.7	1.23	2.2	1.31	2.1	A .12	I 0003
2	42	90	65.0	2.5	1.21	2.4	1.22	2.1	B .16	I 0002
10	78	90	42.0	3.4	1.09	.4	1.06	.2	C .20	I 0010
4	73	90	46.6	2.9	1.04	.2	.97	-.1	D .29	I 0004
7	83	90	35.6	4.1	.98	-.1	.92	-.2	E .23	I 0007
8	72	90	47.4	2.8	.94	-.4	.97	-.1	e .38	I 0008
6	75	90	44.9	3.0	.94	-.3	.82	-.7	d .38	I 0006
5	66	90	51.7	2.6	.92	-.7	.78	-1.3	c .47	I 0005
9	53	90	59.2	2.3	.87	-1.5	.87	-1.3	b .52	I 0009
1	66	90	51.7	2.6	.81	-1.7	.69	-1.9	a .57	I 0001
MEAN	67.	90.	50.0	2.9	1.00	.0	.96	-.1		
S. D.	12.	0.	8.2	.5	.13	1.3	.18	1.3		

A rule of thumb is to limit the range of infit and outfit scores in multiple choice questions to between 0.7 and 1.3. If we are dealing with a high stake test, which is used to make a very important or critical decision about someone's future, for example, we adhere strictly to this rule. Scores 1.31 in Item 3 and 0.69 in Item 1 are beyond this range ;

accordingly, these two items should be removed from the item list.

However, item removal will not be resorted to here, since the present research is intended simply to demonstrate the item banking procedure, and since the number of items is small and the degree of difference from the acceptable range (only on the 0.01 level) does not seem fatal to the analysis. We can leave all items as they are in the list.

In any case, Rasch measurement not only estimates item difficulties and the precision of these estimates but also tests the fit of each item to the construct implied by the set of items. Then, in addition to estimating person measures, it examines the response patterns of persons to determine whether they are responding as expected.

After items are calibrated according to item response theory (IRT) or the Rasch one parameter model in the present research, they can be stored in an item bank according to a common metric of difficulty. This is generally true regardless of the equality of ability or the size of subsequent person samples tested, although an expected minimum number of test takers is needed to make a generalization. The item bank becomes more than just a catalog of items used, with descriptions of their success and failures. It becomes an ever-expanding test which spans the latent ability continuum beyond the measurement needs of any one individual, and it may be accessed to gather items appropriate to any group of persons from the same general population with respect to the ability measured (cf. Henning 1987).

5. Advantages of item banks

Hozayin (2000) says that the main advantage of calibrated item banks is in the ease of test development. A set of items, in the form of a test, may be withdrawn from the bank, and teachers will know how difficult this set of items is for the test takers. The teachers will also know how well these items can discriminate between students who have learned the target content and those who haven't.

Additionally, Hozayin (2000) claims that a second advantage of calibrated item banks is that they can provide the basis for a curriculum map, in which the learning objectives included in the curriculum are ordered by difficulty. This will allow teachers to gain greater insight into the learning process of their students and to confirm that what they think is difficult or easy actually is difficult or easy. It will thus be much easier to chart the progress of individual students over time (cf. Choppin 1979).

Wright and Bell (1984) describe an advantage of item banks from the viewpoint of

students in the following way. A well constructed item bank can provide the basis for designing the best possible test for every purpose. This is because it is not necessary for every student to take the same test in order to be able to compare results. Students can take the selections of bank items most appropriate to their levels of development. The number of items, their level and range of difficulty, and their type and content can be determined for each student individually, without losing the comparability provided by standardized tests. Comparability is maintained because any test formed from bank items, on which a student manifests a valid pattern of performance, is automatically equated, through the calibration of its items onto the bank, to every other test that has been or might be so formed.

Furthermore, Wright and Bell (1984) also point out an advantage of item banks from the viewpoint of teachers. A well-organized item bank enables teachers to construct a wide variety of tests. They need not settle for standard grade level tests or administer the same test to every student in a class or school. They can consider who is to be measured and for what purpose and select items accordingly. They can tailor each test to their immediate educational objectives without losing contact with the common core of bank items. They can write, bank and use new items that reflect their own educational goals while retaining, when their new items fit the bank, the opportunity to make whatever general comparisons they may require.

It is also important to note that because all of the items drawn from a particular bank are calibrated onto one common scale, teachers can compare their test results with one another, even when their tests contain no common items (Wright and Bell 1984). This opportunity to compare results quantitatively enables teachers to examine how the same topic is learned by different students working with different teachers and hence to evaluate alternative teaching strategies. With common curriculum strands as the frames of reference, it becomes possible to recognize subtle differences in the way school subjects are mastered. The investigation of which teaching methods are most effective in which circumstances can become an ongoing, routine part of the educational process. In other words, tests constructed from item banks can promote an exchange of ideas, not only about assessment, but also about curricula (Wright and Bell 1984).

6. Limitations of item banks

As with any approach to educational measurement, there are limitations on item

banks. Using an item bank will not eliminate the need for test developers to evaluate the quality of the items stored in the bank. In addition, the test developers must be sure that the content tested by the item reflects the target content (Hozayin 2000).

Furthermore, Choppin (1979) says that it is important to realize that item banking is not the final solution to all the problems posed by educational assessment. No item bank can be better than the material that is put into it, and users of assessment materials will continue to carry responsibility for ensuring that their tests are fair, appropriate, reliable and valid. An item bank should be a living thing with test materials being added and the classification system updated as new developments occur either in our understanding of the subject matter or in teaching practices (Choppin 1979).

7. Conclusions

Item response theory facilitates item banking by allowing all of the items to be calibrated and positioned on the same latent continuum by means of a common metric. Also, it permits additional items to be added subsequently without the need to locate and retest the original sample of examinees. Furthermore, an item bank permits the construction of tests of known reliability and validity based on appropriate selection of item subsets from the bank without further need for trial in the field (Henning, 1987).

Hozayin (2000) stresses the point that a carefully developed item bank may serve as the basis for adaptive testing, which is usually called computer adaptive testing (CAT), (since adaptive tests are almost always delivered on a computer). This allows item selection to match the specific ability level of the individual student who is taking the test.

Finally, we have learned how item calibrations and person measurement are conducted using the Rasch model for item banking. The idea of item banking, along with improvements in computer technology, will lead to a new approach to language test development and use, even though there may be hurdles to be cleared in the process.

Note: This research was supported in part by Tokyo Keizai University under Research Grant CPU04-00.

Acknowledgment

I am grateful to Dr. Benjamin D. Wright and Dr. John M. Linacre for their invaluable comments.

Bibliography

Beeston, S. (2000). UCLES Research Notes 2. University of Cambridge Local Examinations Syndicate.

Bode, R.K. & Wright, B.D. (1999). Rasch measurement in higher education. Higher Education : Handbook of Theory and Research, vol. XIV. (pp. 287-316).

Choppin, B. (1979). Testing the questions—the Rasch model and item banking. In M. St. J. Raggett, C. Tutt, P. Raggett. (Eds.). Assessment and Testing of Reading : Problems and Practice. London : Ward Lock Educational.

Davies, A, A. Brown, C. Elder, K. Hill, T. Lumley and T. McNamara. (1999). Studies in Language Testing 7 : Dictionary of language testing. Cambridge UK : Cambridge University Press.

Gronlund, N. (1998). Assessment of Student Achievement. 6th Edition. Needham Heights, MA : Allyn and Bacon.

Henning, G. (1987). A Guide to Language Testing. New York, NY : Newbury House Publishers.

Hozayin, R. (2000). Item Banks : Definition and Development. Paper presented at the Sixth EFL Skills Conference at the American University in Cairo, January 25th—27th, 2000.

Linacre, John M. (1989, 1994). Many-Facet Rasch Measurement. Chicago, IL : MESA Press.

Linacre, John M. & Wright, B. (1998, 2001). A User's Guide to Bigsteps/Winsteps : Rasch-Model Computer Program. Chicago, IL : MESA Press.

Rudner, L. (1998). Item banking. ERIC/AE Digest Series EDO-TM-98-05.

Wright, B. & Bell, S. (1984). Item banks. What, why, how. Journal of Educational Measurement. 21 (4), 331-345 Winter.

Wright, B. & Stone, M. (1979). Best Test Design : Rasch Measurement. Chicago, IL : MESA Press.

Appendix 1

Subjects

The subjects in this research were 105 Japanese university students majoring in business administration. Most of them were from eighteen to twenty-one years old of age.

Items

The 10 items were based on the TOEFL listening comprehension dialogue test format. In the test, students listened to a dialogue (which is usually between a man and a woman) followed by-a narrator's question. Then, they had to choose the correct answer from among the four multiple-choice options for ten items written in English.

The topics for the 10 items were as follows :

1) a talk at a parking lot (item 1)
2) a talk about a vacation plan (item 2)
3) a talk about a party preparation (item 3)
4) a talk about a telephone message (item 4)
5) a talk about the weekend camping (item 5)

6) a talk about a rock concert (item 6)
7) a talk about a dinner party (item 7)
8) a talk about a car accident (item 8)
9) a talk about an election (item 9)
10) a talk at a shop (item 10)

Appendix 2

INFIT is an information-weighted fit statistic, which is more sensitive to unexpected behavior affecting responses to items near the person's ability level.

MNSQ is the mean-square infit statistic with expectation 1.

ZSTD is the infit mean-square fit statistic standardized to approximate a theoretical mean 0 and variance 1 distribution.

OUTFIT is an outlier-sensitive fit statistic, which is more sensitive to unexpected behavior by persons on items far from the person's ability level.

MNSQ is the mean-square outfit statistic with expectation 1.

ZSTD is the outfit mean-square fit statistic standardized to approximate a theoretical mean 0 and variance 1 distribution.

REALSE

SE is the standard error computed over the persons or over the items.

REALSE is computed on the basis that misfit in the data is due to departures in the data from model specifications. (cf. Linacre and Wright, 1998).